Convergence rates of posterior distributions for infinite-dimensional exponential families

C. Scricciolo

2001.20

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Dicembre 2001

CONVERGENCE RATES OF POSTERIOR DISTRIBUTIONS FOR INFINITE-DIMENSIONAL EXPONENTIAL FAMILIES

CATIA SCRICCIOLO

ABSTRACT. In this paper some prior distributions for densities in infinite-dimensional exponential families, whose logarithm has a prescribed degree of smoothness, are designed so that the corresponding posteriors converge at the optimal rate. The derived Bayes' density estimator attains the minimax rate under the Hellinger loss.

1. INTRODUCTION

In this paper we address the problem of determining rates of convergence for posterior distributions on an infinite-dimensional exponential family of densities whose logarithm is supposed to have a certain degree of smoothness. The final goal is to construct priors achieving the optimal rate in the minimax sense for point estimators. The mathematical formulation of the problem follows.

Suppose that $X_1, \ldots, X_n$ is a random sample of n i.i.d. observations drawn from an unknown distribution P_0 having density f_0 w.r.t. a finite measure λ. Suppose P_0 to be supported on a bounded interval, which, for convenience, can be taken to be $[0, 1]$ with the Lebesgue measure. The generic density of the model is assumed to be of the form

$$f_\theta(x) = \exp\{\theta(x) - \psi(\theta)\}, \qquad x \in [0, 1], \tag{1.1}$$

where the parameter θ belongs to a subset $\Theta_p(B)$ of $L_2[0, 1]$, the separable Hilbert space of all square-integrable functions on $[0, 1]$, endowed with the inner product $\langle \theta, \eta \rangle = \int_0^1 \theta(t)\eta(t)\,\mathrm{d}t$, and $\psi(\theta)$ is the normalizing constant, i.e.,

$$\psi(\theta) = \ln\left(\int_0^1 \exp\{\theta(t)\}\,\mathrm{d}t\right), \qquad \theta \in \Theta_p(B).$$

For a fixed integer $p \geq 1$ and a constant $B > 0$, the parameter space $\Theta_p(B)$ is defined as

$$\Theta_p(B) = \left\{\theta \in L_2[0, 1] : \theta(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x), \quad \sum_{j=1}^{\infty} \theta_j^2 j^{2p} \leq B\right\},$$

where $\sum_{j=1}^{\infty} \theta_j \phi_j$ is the Fourier representation of θ w.r.t. to the orthonormal trigonometric basis $\{\phi_j\}_{j=1}^{\infty}$ for $L_2[0, 1]$. The basis functions ϕ_j and the corresponding coefficients θ_j are defined as follows:

$$\phi_1(x) \equiv 1, \quad \begin{cases} \phi_2(x) = \sqrt{2}\cos(2\pi x) \\ \phi_3(x) = \sqrt{2}\sin(2\pi x) \end{cases}, \ldots, \begin{cases} \phi_{2k}(x) = \sqrt{2}\cos(2k\pi x) \\ \phi_{2k+1}(x) = \sqrt{2}\sin(2k\pi x) \end{cases}, \ldots, k \in \mathbb{N},$$

and

$$\theta_1 = \langle \theta, \phi_1 \rangle, \quad \begin{cases} \theta_2 = \langle \theta, \phi_2 \rangle \\ \theta_3 = \langle \theta, \phi_3 \rangle \end{cases}, \ldots, \begin{cases} \theta_{2k} = \langle \theta, \phi_{2k} \rangle \\ \theta_{2k+1} = \langle \theta, \phi_{2k+1} \rangle \end{cases}, \ldots, k \in \mathbb{N}.$$

Any $\theta \in \Theta_p(B)$ can be represented as

$$\theta(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x), \qquad x \in [0, 1],$$

and is *uniquely* determined by its Fourier coefficients. It is worthwhile spending few words on the meaning of the previous equality. In the first place, it means that the sequence of partial sums $\{\sum_{j=1}^{m} \theta_j \phi_j\}_{m=1}^{\infty}$ converges to θ in the L_2-norm, i.e.,

$$\lim_{m \to \infty} \left\| \theta - \sum_{j=1}^{m} \theta_j \phi_j \right\|_2 = 0.$$

In other terms, $\sum_{j=1}^{\infty} \theta_j \phi_j$ is the expansion of θ w.r.t. the Schauder basis $\{\phi_j\}_{j=1}^{\infty}$ and it is fairly natural to write $\theta = \sum_{j=1}^{\infty} \theta_j \phi_j$. In the second place, it means that convergence of the trigonometric Fourier series to θ is point-wise.

The density in (1.1) can be rewritten as

$$f_\theta(x) = \frac{\exp\left\{\sum_{j=1}^{\infty} \theta_j \phi_j(x)\right\}}{\int_0^1 \exp\{\sum_{j=1}^{\infty} \theta_j \phi_j(t)\}\, dt}, \qquad x \in [0, 1],$$

and the model $\mathcal{F} = \{P_\theta, \theta \in \Theta_p(B)\}$ is identifiable. The sampling distribution P_0 corresponds to P_{θ_0}, for some $\theta_0 \in \Theta_p(B)$. Let P_0^n stand for the n-fold product measure of P_0.

Some facts concerning the parameter space are highlighted. The assumption that the log-density belongs to $\Theta_p(B)$ is a smoothness condition. It implies that the log-density has a degree of smoothness at least p and that the pth derivative is bounded in the L_2-norm: it is easy to see that $\|\theta^{(p)}\|_2^2 = \sum_{k=1}^{\infty} (2k\pi)^{2p} (\theta_{2k}^2 + \theta_{2k+1}^2)$. By requiring that $\theta \in \Theta_p(B)$, the density f_θ is forced to be positive and uniformly bounded. As the basis functions ϕ_j are uniformly bounded on $[0, 1]$, there exists a real number $c_1 > 0$, depending on both p and B, such that

$$\sup_{\theta \in \Theta_p(B)} \|\theta\|_\infty \leq c_1, \tag{1.2}$$

where $\|\theta\|_\infty$ stands for the sup-norm of θ, i.e., $\|\theta\|_\infty = \sup_{x \in [0,1]} |\theta(x)|$. Consequently, also $|\psi(\theta)| \leq c_1$ uniformly on $\Theta_p(B)$ and $\sup_{\theta \in \Theta_p(B)} \|f_\theta\|_\infty \leq e^{2c_1}$.

Let μ be a prior on $\Theta_p(B)$. Let ε_n be a positive sequence tending to zero. For any positive sequence K_n such that $K_n \to \infty$ and $K_n\varepsilon_n \to 0$, as $n \to \infty$, let $A_n = \{\theta \in \Theta_p(B) : \|\theta - \theta_0\|_2 < K_n\varepsilon_n\}$ be a L_2-shrinking neighborhood of θ_0. We say that the (point-wise) rate of convergence of the posterior is ε_n if, for each sequence K_n as before, $\mu(A_n^c|X^n)$ tends to zero in probability or almost surely when sampling from P_0. Since μ induces a prior on $\mathcal{F}$, say Π, through the map $\theta \mapsto P_\theta$, the same definition applies to the posterior probability of a Hellinger neighborhood $H_{\varepsilon_n}^c$ of P_0. Recall that the Hellinger distance between P_0 and P_θ is defined as

$$d_H(P_0, P_\theta) = \left\{ \int_0^1 \left(\sqrt{f_0(x)} - \sqrt{f_\theta(x)} \right)^2 \mathrm{d}x \right\}^{1/2}.$$

For the sake of brevity, we shall henceforth omit the argument of densities when this will cause no ambiguity. General results on rates of convergence of posteriors are available from the works of Ghosal *et al.* (2000) and Shen *et al.* (2001). In order to quote a theorem for later reference, we need some definitions. For any $t > 0$, let $M(t) = \{P_\theta \in \mathcal{F} : \max\{K(P_0\|P_\theta), D(P_0, P_\theta)\} < t\}$, where $D(P_0, P_\theta) = \int_0^1 f_0 \left(\ln(f_0/f_\theta)\right)^2 - \left(\int_0^1 f_0 \ln(f_0/f_\theta)\right)^2$ and $K(P_0\|P_\theta)$ is the Kullback-Leibler divergence:

$$K(P_0\|P_\theta) = \begin{cases} \int_0^1 f_0 \ln(f_0/f_\theta), & \text{if } P_0 \ll P_\theta, \\ \infty, & \text{if } P_0 \not\ll P_\theta. \end{cases}$$

In the sequel, for a positive sequence t_n, let M_n stand for $M(t_n)$. A further convention. When writing $a_n \gtrsim b_n$, we mean that the inequality $a_n \geq cb_n$ holds for all sufficiently large n, with $c > 0$ a constant that is fixed throughout. When determining rates of convergence, constants are not relevant. An analogue meaning has $a_n \lesssim b_n$. If both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold true, then we simply write $a_n \asymp b_n$.

Theorem 1.1. [Shen and Wasserman (2001)]. *Suppose that for positive sequences r_n and t_n such that $\varepsilon_n = \max\{r_n, t_n^{1/2}\} \to 0$ and $n\min\{r_n^2, t_n\} \to \infty$, as $n \to \infty$, constants $a > 0$ and $c = (2/3)^{5/2}/512$, it is*

$$\int_{r_n^2/2^8}^{\sqrt{2}r_n} \sqrt{H_{[]}(u/a, \mathcal{F})}\,\mathrm{d}u \leq c\sqrt{n}r_n^2, \tag{1.3}$$

$$\pi(M_n) \gtrsim e^{-2nt_n}, \tag{1.4}$$

where $H_{[]}(\cdot, \mathcal{F})$ denotes the Hellinger bracketing metric entropy of $\mathcal{F}$. Then, for a sufficiently large constant $K > 0$, $\pi(H_{\varepsilon_n}^c|x^n) \preceq \exp\left\{-c_1K^2n\varepsilon_n^2/2\right\}$ on a Borel set of P_0^n-probability tending to one.

A prior probability measure can be induced on $\Theta_p(B)$ by assigning a distribution to the Fourier coefficients of the orthonormal series expansion. It is quite natural to take the θ_j's to be independent normals with zero mean and variance τ_j^2. In principle, a Gaussian process could be directly defined on $\Theta_p(B)$, however, the first

approach is preferred to the second one for mathematical convenience. This model is closely related to the logistic normal process of Lenk (1988, 1991). The prior remains determined by the choice of the τ_j's. The problem is to elicit the prior variances in such a way that the posterior achieves a desired rate uniformly on $\Theta_p(B)$. The target rate is optimal in the sense made precise soon after.

Let $\hat{\theta}$ be an estimator of $\theta \in \Theta_p(B)$. If the minimax risk $\inf_{\hat{\theta}} \sup_{\theta \in \Theta_p(B)} R(\hat{\theta}, \theta)$, with $R(\hat{\theta}, \theta) = \mathrm{E}_\theta[\|\hat{\theta} - \theta\|_2^2]$, is considered, then the optimal rate of convergence is known to be $r_n^2 \propto n^{-2p/(2p+1)}$, i.e.,

$$0 < \lim_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta \in \Theta_p(B)} n^{2p/(2p+1)} R(\hat{\theta}, \theta) < \infty.$$

This is a long history result first established by Ibragimov *et al.* (1977, 1981), see also Pinsker (1980). This entails that, for any prior supported on $\Theta_p(B)$, the corresponding posterior cannot concentrate on L_2-balls of θ_0 at a rate faster than r_n. Since Hellinger neighborhoods of P_0 correspond to L_2-neighborhoods of θ_0, if a prior on $\Theta_p(B)$ can be defined so that the posterior converges at the optimal rate, the induced posterior on $\mathcal{F}$ will converge at the same rate.

Infinite-dimensional exponential families have been considered in different settings. Consistency issues have been studied by Barron (1988), Barron *et al.* (1999) and Walker *et al.* (2001). Crain (1974, 1976) studied the problem of estimation for densities from exponential families generated by Legendre polynomials; the same model has been effectively used by Verdinelli *et al.* (1998) for Bayesian goodness-of-fit testing. As far as the author is aware, rates of convergence for posterior distributions have not been studied before, at least in the present framework.

We briefly sketch the program. In order to appeal to Theorem 1.1, in the next section, we derive basic tools allowing us to show that the bracketing integral equation provides exactly the rate $r_n \propto n^{-p/(2p+1)}$. Since Kullback-Leibler type neighborhoods M_n of P_0 translate into L_2-neighborhoods of θ_0, requirement (1.4) is met if a prior for θ can be defined that assigns probability mass not less than $e^{-Cnr_n^2}$, for some constant $C > 0$, to L_2-neighborhoods of θ_0.

The paper is organized as follows. In Section 3, we start considering a sequence of independent normal priors and show that, unless the prior variances are chosen to die off rapidly enough, the posterior does not converge at the intended rate. As shown in Sections 4 and 5, the problem can be fixed by using either a sample size-dependent prior or a "sieve" prior.

2. Auxiliary lemmas

The densities f_θ are uniformly bounded and uniformly bounded away from zero. Consequently, the Hellinger distance, the Kullback-Leibler divergence and the χ^2-divergence are equivalent: each one is upper and lower bounded by multiples of the others. Moreover, as it is herein shown, for any pair P_{θ_0}, P_θ, they are upper and lower bounded by multiples of the L_2-distance between θ_0 and θ. This descends from the fact that probability measures in $\mathcal{F}$ are compactly supported and θ is uniformly bounded in the sup-norm. The starting point is the following lemma due to Barron *et al.* (1991, pages 1355-1356, Lemma 1). By the notation $\|p\|_{L_\infty}$, the L_∞-norm w.r.t. a measure λ is meant, i.e.,

$$\|p\|_{L_\infty} = \operatorname*{ess\,sup}_{[\lambda]} |p| = \inf\{v : \lambda(\{x : |p(x)| > v\}) = 0\}.$$

Lemma 2.1. [BARRON AND SHEU (1991)]. *Let P and Q be probability measures on a σ-finite measure space $(\mathcal{X}, \mathcal{B}, \lambda)$ and let p and q denote versions of their Radon-Nikodym derivatives w.r.t. λ. If $\|\ln p/q\|_{L_\infty} < \infty$, then*

$$K(P\|Q) \geq \frac{1}{2} e^{-\|\ln p/q\|_{L_\infty}} \int_{\mathcal{X}} p \left(\ln \frac{p}{q}\right)^2 d\lambda \tag{2.1}$$

and

$$K(P\|Q) \leq \frac{1}{2} e^{\|\ln p/q - c\|_{L_\infty}} \int_{\mathcal{X}} p \left(\ln \frac{p}{q} - c\right)^2 d\lambda, \tag{2.2}$$

where c is any constant.

In the sequel, we use the notation $a \lesssim b$ to mean that $a \leq cb$, for a constant c that is fixed throughout. Analogously, we write $a \gtrsim b$ if the reverse inequality holds.

Lemma 2.2. *For any pair $\theta_0, \theta \in \Theta_p(B)$,*

$$K(P_{\theta_0}\|P_\theta) \lesssim \|\theta - \theta_0\|_2^2, \tag{2.3}$$

$$\mathrm{E}_{\theta_0}\left[\ln \frac{f_{\theta_0}(X_1)}{f_\theta(X_1)}\right]^2 \lesssim \|\theta - \theta_0\|_2^2. \tag{2.4}$$

Proof. In order to appeal to Lemma 2.1, we need to check that $\|\ln f_{\theta_0}/f_\theta\|_{L_\infty} < \infty$. Note that

$$|\psi(\theta) - \psi(\theta_0)| \leq \|\theta - \theta_0\|_\infty. \tag{2.5}$$

Now, in force of (2.5), for any $x \in [0, 1]$,

$$\left|\ln \frac{f_{\theta_0}(x)}{f_\theta(x)}\right| = |[\theta_0(x) - \theta(x)] + [\psi(\theta) - \psi(\theta_0)]|$$

$$\leq 2\|\theta - \theta_0\|_\infty \leq 4c_1, \tag{2.6}$$

hence,

$$\|\ln f_{\theta_0}/f_\theta\|_\infty \leq 4c_1. \tag{2.7}$$

Now, we prove the bound in (2.3). To this aim, consider (2.2) with $c = \psi(\theta) - \psi(\theta_0)$. It results that

$$K(P_{\theta_0}||P_\theta) \leq \frac{1}{2} e^{\|\theta_0 - \theta\|_{L_\infty}} \int_0^1 [\theta_0(x) - \theta(x)]^2 f_{\theta_0}(x)\,\mathrm{d}x$$

$$\leq \frac{1}{2} e^{\|\theta_0 - \theta\|_\infty} \|f_{\theta_0}\|_\infty \|\theta - \theta_0\|_2^2 \leq \frac{1}{2} e^{4c_1} \|\theta - \theta_0\|_2^2 . \tag{2.8}$$

Lastly, we show (2.4). In virtue of (2.1), by combining (2.7) and (2.8), we have that

$$\mathrm{E}_{\theta_0}\left[\ln \frac{f_{\theta_0}(X_1)}{f_\theta(X_1)}\right]^2 \leq 2e^{\|\ln f_{\theta_0}/f_\theta\|_{L_\infty}} K(P_{\theta_0}||P_\theta) \leq e^{8c_1} \|\theta - \theta_0\|_2^2 .$$

□

We recall that the χ^2-divergence of any pair of probability measures P and Q is defined as $\chi^2(P||Q) = \int_{\{pq>0\}} (p^2/q)\,\mathrm{d}\lambda - 1 + \infty P(\{q = 0\})$.

Lemma 2.3. *For any pair θ_0, $\theta \in \Theta_p(B)$,*

$$\|\theta - \theta_0\|_2^2 \lesssim d_H(P_{\theta_0}, P_\theta)^2 \lesssim \chi^2(P_{\theta_0}||P_\theta) \lesssim \|\theta - \theta_0\|_2^2 , \tag{2.9}$$

and so

$$d_H(P_{\theta_0}, P_\theta) \lesssim \|\theta - \theta_0\|_\infty . \tag{2.10}$$

Proof. First the lower bound in (2.9) is shown. The following chain of inequalities is seen to hold true:

$$\begin{aligned} d_H(P_{\theta_0}, P_\theta)^2 &\geq \frac{1}{2}\left(\left\|\frac{f_{\theta_0}}{f_\theta}\right\|_\infty\right)^{-1} K(P_{\theta_0}||P_\theta) \\ &\geq \frac{1}{4}\left(\left\|\frac{f_{\theta_0}}{f_\theta}\right\|_\infty\right)^{-1} e^{-\|\ln f_{\theta_0}/f_\theta\|_{L_\infty}} \int_0^1 \left[\ln \frac{f_{\theta_0}(x)}{f_\theta(x)}\right]^2 f_{\theta_0}(x)\,\mathrm{d}x \\ &\geq \frac{1}{4} e^{-2c_1}\left(\left\|\frac{f_{\theta_0}}{f_\theta}\right\|_\infty\right)^{-1} e^{-\|\ln f_{\theta_0}/f_\theta\|_{L_\infty}} \|\ln f_{\theta_0} - \ln f_\theta\|_2^2 \\ &\geq \frac{1}{4} e^{-10c_1} \|\theta - \theta_0\|_2^2 . \end{aligned}$$

The first line descends from (7.6) of Lemma 5 in Birgé *et al.* (1998, pages 361-362), the second line follows from (2.1) and the last one from (2.6) and the inequality

$\|\ln f_{\theta_0} - \ln f_\theta\|_2^2 \geq [\psi(\theta_0) - \psi(\theta)]^2 + \|\theta_0 - \theta\|_2^2$, which is verified by straightforward calculations.

The upper bound in (2.9) is deduced from (2.3), recalling that, for any pair of probability measures P and Q, $d_H(P, Q)^2 \leq K(P\|Q)$ and, if $\|f_P/f_Q\|_\infty < \infty$,

$$d_H(P, Q)^2 \leq \chi^2(P\|Q) \leq 4 \left\| \frac{f_P}{f_Q} \right\|_\infty d_H(P, Q)^2.$$

□

We are now in a position to show that the bracketing integral equation is satisfied with $r_n \propto n^{-p/(2p+1)}$. In force of (2.10), the inequality

$$H_{[]}(u, \mathcal{F}) \leq c \ln N(u, \Theta_p(B), \|\cdot\|_\infty)$$

holds true for some constant $c > 0$ and all $u > 0$, where $\ln N(u, \Theta_p(B), \|\cdot\|_\infty)$ denotes the sup-norm metric entropy of $\Theta_p(B)$. From Birman and Solomjak (1967), it is known that, for each $u > 0$, $\ln N(u, \Theta_p(B), \|\cdot\|_\infty) \leq Au^{-1/p}$, for some $A > 0$. Hence, $H_{[]}(u, \mathcal{F}) \leq A'u^{-1/p}$ and the equation

$$\int_{r_n^2/2^8}^{\sqrt{2}r_n} \sqrt{H_{[]}(u/a, \mathcal{F})}\, du = c\sqrt{n}r_n^2$$

gives the solution $r_n \propto n^{-p/(2p+1)}$. We do not delve into details, we just observe that $r_n^{-1/2p} = \sqrt{n}r_n$.

3. Infinite independent normal priors

We start with a sequence of independent, normally distributed Fourier coefficients. The assumption of independence seems reasonable in view of the orthogonality of the basis functions. We assume that each $\theta_j \sim N(0, \tau_j^2)$, with $\tau_j = j^{-q}$, for a fixed $q \geq p + 1/2$. The joint prior density of the sequence $\{\theta_j\}_{j=1}^\infty$ is

$$\pi(\theta_1, \theta_2, \ldots) = \prod_{j=1}^\infty \pi_j(\theta_j) = \prod_{j=1}^\infty \frac{1}{\sqrt{2\pi j^{-2q}}} \exp\left\{ -\frac{\theta_j^2}{2j^{-2q}} \right\}. \tag{3.1}$$

Since $\sum_{j=1}^\infty \tau_j^2 \phi_j(x)^2 < \infty$ for all $x \in [0, 1]$, the series $\sum_{j=1}^\infty \theta_j \phi_j(x)$ converges a.s. to a Gaussian process, with covariance function $\sigma(x, y) = \sum_{j=1}^\infty \tau_j^2 \phi_j(x)\phi_j(y)$. In other terms, $\sum_{j=1}^\infty \theta_j \phi_j(x)$ defines a random function living in $L_2[0, 1]$, while it should live in $\Theta_p(B)$. If $q = p + 1/2$, then the prior gives zero mass to $\Theta_p(B)$, whereas, if $q > p + 1/2$, then the prior puts positive mass on $\Theta_p(B)$ and it can be truncated to sit on it. A detailed explanation of this phenomenon can be found in Zhao (2000, page 541). In the sequel, with abuse of notation, we shall use π to denote both the prior for the coefficients and the induced prior for the whole function θ, the right meaning being clear form context. Let μ denote the restriction of π to $\Theta_p(B)$: for each Borel set $A \subseteq L_2[0, 1]$, $\mu(A)$ is defined as

$$\mu(A) = \frac{\pi(A \cap \Theta_p(B))}{\pi(\Theta_p(B))}.$$

Let Π denote the prior on $\mathcal{F}$ induced by μ. The posterior led by Π should hopefully converge at the desired rate. We shall prove the main result of this section by appealing to the following result, a specialization of Lemma 5 in Shen *et al.* (2001, page 711-712).

Lemma 3.1. [Shen and Wasserman (2001)]. *Let $\{Z_j\}_{j=1}^{\infty}$ be a sequence of i.i.d. r.v.'s, with $Z_j \sim N(0,1)$. Let $a_j = j^{-d}$, with $d \geq p$, for a fixed integer $p \geq 1$. Let $\delta > 0$ and let N be the smallest integer such that*

$$\sum_{j=N+1}^{\infty} j^{-2d} \leq \delta^2/4. \tag{3.2}$$

Then, for a positive constant c',

$$Pr\left(\sum_{j=1}^{\infty}(a_j Z_j)^2 \leq \delta^2\right) \geq c' \exp\left\{-\left(d+\frac{1}{2}\right)N\right\}.$$

Theorem 3.2. *Let Π be the prior measure on $\mathcal{F}$ induced by the restriction of (3.1) to $\Theta_p(B)$. For each $P_0 \in \mathcal{F}$, if $q > 2p + 1/2$, then $\pi(H_{\varepsilon_n}^c | X^n)$ tends to zero in P_0^n-probability at the rate $\varepsilon_n \propto n^{-p/(2p+1)}$.*

Proof. In view of Theorem 1.1, the assertion follows if $\Pi(M_n) \gtrsim e^{-Cnr_n^2}$, for some $C > 0$. In force of the chain of set inclusions

$$\begin{aligned}\left\{\theta \in \Theta_p(B) : \|\theta - \theta_0\|_2^2 < cr_n^2\right\} &\\ \subseteq \left\{\theta \in \Theta_p(B) : \max\left\{K(P_0\|P_\theta),\, \mathrm{E}_0\left[\ln(f_0(X_1)/f_\theta(X_1))^2\right]\right\} < r_n^2\right\} &\\ \subseteq \left\{\theta \in \Theta_p(B) : \max\left\{K(P_0\|P_\theta),\, D(P_0, P_\theta)\right\} < r_n^2\right\} &\end{aligned} \tag{3.3}$$

which holds for a universal constant $c > 0$ in virtue of (2.3) and (2.4), the subset M_n of $\mathcal{F}$ corresponds to a subset of $\Theta_p(B)$, whose prior probability can be lower bounded by bounding below the probability of the set in (3.3). For n large enough, it is

$$\begin{aligned}\mu\left(\left\{\theta \in \Theta_p(B) : \|\theta - \theta_0\|_2^2 < cr_n^2\right\}\right) &\geq \pi\left(\left\{\theta : \|\theta - \theta_0\|_2^2 < cr_n^2,\ \sum_{j=1}^{\infty}\theta_j^2 j^{2p} \leq B\right\}\right) \\ &= \pi\left(\left\{\theta : \sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 < cr_n^2,\ \sum_{j=1}^{\infty}\theta_j^2 j^{2p} \leq B\right\}\right) \\ &\geq \pi\left(\left\{\theta : \sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 j^{2p} < cr_n^2\right\}\right), \qquad (3.4)\end{aligned}$$

where the last inequality follows from the inclusion

$$\left\{\theta : \sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 j^{2p} < cr_n^2\right\} \subseteq \left\{\theta : \sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 < cr_n^2, \quad \sum_{j=1}^{\infty}\theta_j^2 j^{2p} \leq B\right\},$$

which holds true for sufficiently large n. To see it, rename the set on the left-hand side to U_n and that on the right-hand side to V_n. Note that, if $\theta \in U_n$, then $\sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 \leq \sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 j^{2p} \leq cr_n^2$, hence $\sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 \leq cr_n^2$. It has to be shown that θ satisfies also the constraint $\sum_{j=1}^{\infty}\theta_j^2 j^{2p} \leq B$. Since $\theta_0 \in \Theta_p(B)$, there exists $\beta_0 \equiv \beta(\theta_0) \geq 0$ such that $\sum_{j=1}^{\infty}\theta_{0j}^2 j^{2p} = B - \beta_0$. Reasoning as in Shen *et al.* (2001, page 708), it can be shown that, if $\sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 j^{2p} < cr_n^2$, then for n large enough so that cr_n^2 is sufficiently small,

$$\begin{aligned}\sum_{j=1}^{\infty}\theta_j^2 j^{2p} &\leq \sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 j^{2p} + \sum_{j=1}^{\infty}\theta_{0j}^2 j^{2p} + 2\sqrt{\sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 j^{2p}}\sqrt{\sum_{j=1}^{\infty}\theta_{0j}^2 j^{2p}} \\ &< cr_n^2 + (B - \beta_0) + 2\sqrt{cr_n^2(B - \beta_0)} \\ &= B - \frac{\beta_0}{2},\end{aligned}$$

hence, $\sum_{j=1}^{\infty}\theta_j^2 j^{2p} \leq B$ and $\theta \in V_n$.

In order to lower bound (3.4), we appeal to Lemma 3.1, with $d = q - p$, $\delta^2 = cr_n^2/2$ and $N = O(r_n^{-1/(d-1/2)})$, as determined from (3.2). Thus, for a suitable constant $B > 0$, setting the positions $\Delta = q - (p + 1/2)$ and $t_n = n^{-[(2p+1-p/\Delta)/(2p+1)]}$, for large n, it turns out to be

$$\begin{aligned}\pi\left(\left\{\theta : \sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 j^{2p} < cr_n^2\right\}\right) &\gtrsim \exp\left\{-Bn^{\frac{p/(d-1/2)}{2p+1}}\right\} \\ &= \exp\left\{-Bn^{\frac{p/(q-p-1/2)}{2p+1}}\right\} \\ &= \exp\left\{-Bn^{\frac{p/\Delta}{2p+1}}\right\} \\ &= \exp\left\{-Bnt_n\right\}.\end{aligned}$$

Depending on the possible values of the ratio p/Δ, two cases can be distinguished. For $q \geq 2p + 1/2$, it is $t_n \leq r_n^2$, hence, the posterior converges at a rate which is a multiple of $n^{-p/(2p+1)}$. For $q \in (p + 1/2 + p/(2p + 1), 2p + 1/2)$, it is $t_n > r_n^2$ and the posterior converges at a suboptimal rate. □

The result can indeed be enhanced to hold along almost all sample paths when sampling from P_0.

Theorem 3.3. *Under the same assumptions of Theorem 3.2, for each $P_0 \in \mathcal{F}$, the posterior converges almost surely at the rate $\varepsilon_n \propto n^{-p/(2p+1)}$ relative to the Hellinger distance.*

Proof. An almost-sure version of Theorem 1.1 can be derived by appealing to Lemma 2 instead of Lemma 1 by Shen *et al.* (2001, page 691). If the χ^2-divergence is considered (this corresponds to take $\alpha = 1$), it must be shown that the prior assigns not exponentially small prior probability to neighborhoods $S_n = \{P_\theta : \chi^2(P_0||P_\theta) < t_n\}$. In force of (2.9), χ^2-neighborhoods of P_0 correspond to L_2-neighborhoods of θ_0. Reasoning as in the preceding theorem, it is seen that $\Pi(S_n) \gtrsim e^{-Cnr_n^2}$, for $C > 0$, and the assertion follows. □

The previous findings state that, in order to ensure convergence at the optimal rate, the prior variances should be suitably chosen so as to die off rapidly enough. We can attempt an intuitive explanation of the implications entailed by this choice. Since the sequence θ_j converges to zero, by taking the prior variances of the last terms very small, we are actually forcing the θ_j's to be stuck to small values, that is, a low weight is being assigned to the superior armonics in the Fourier series expansion of θ. In other words, we are implicitly assuming that, with high probability, θ is well-approximated by its projections on finite-dimensional subspaces of $\Theta_p(B)$.

Now, we consider the Bayes' density estimator and draw conclusions on its rate of convergence. Keeping in mind the connection of the present model with a logistic process, the Bayes' estimator can be given an explicit expression:

$$\widehat{f}_n(x) = \exp\left\{\frac{\sigma(x,\, x)}{2}\right\} \frac{\exp\left\{\sum_{i=1}^n \sigma(x,\, x_i)\right\}}{\left(\int_0^1 \exp\{\theta(t)\}\, \mathrm{d}t\right)^{n+1}}, \qquad x \in [0,\, 1],$$

which, however, involves an integral difficult to compute. The problem can be circumvented by resorting to approximations as in Lenk (1988, pages 512-513).

Corollary 3.1. *Under the hypotheses of Theorem 3.2, the Bayes' density estimator $\widehat{f}_n$ achieves the minimax rate $n^{-p/(2p+1)}$ under the Hellinger loss.*

Proof. On the one hand, in force of Theorem 5 of Shen *et al.* (2001, page 694),

$$d_H(f_0,\, \hat{f}_n)^2 \preceq n^{-2p/(2p+1)} + 2e^{-cn^{1/(2p+1)}/2}$$

for some $c > 0$, on a set of P_0^n-probability tending to one. Hence,

$$\mathrm{E}_0^n[d_H(f_0,\, \hat{f}_n)^2] \asymp \sup_{f_0 \in \mathcal{F}} \mathrm{E}_0^n[d_H(f_0,\, \hat{f}_n)^2] \preceq n^{-2p/(2p+1)}. \tag{3.5}$$

On the other hand, in virtue of Corollary 1 by Yang *et al.* (1999, page 1574), it is

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}} \mathrm{E}_0^n[\|f_0 - \hat{f}\|_2^2] \asymp \inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}} \mathrm{E}_0^n[d_H(f_0, \hat{f})^2] \asymp n^{-2p/(2p+1)}.$$

It follows that

$$n^{-2p/(2p+1)} \preceq \inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}} \mathrm{E}_0^n[d_H(f_0, \hat{f})^2] \preceq \sup_{f_0 \in \mathcal{F}} \mathrm{E}_0^n[d_H(f_0, \hat{f}_n)^2]. \tag{3.6}$$

By combining (3.5) and (3.6), we have that

$$\sup_{f_0 \in \mathcal{F}} \mathrm{E}_0^n[d_H(f_0, \hat{f}_n)^2] \asymp n^{-2p/(2p+1)}.$$

□

4. Convergence rate adopting a sample size-dependent prior

It is herein shown that one possible way of fixing the problem is to use a carefully chosen sample-size dependent prior. For a fixed $p \geq 1$, let $q = p + 1/2$ and $k_n = O(r_n^{-1/p})$, i.e., $k_n \propto n^{1/(2p+1)}$. For simplicity of notation, we drop the subscript n. Let Θ_k be the subset of $\Theta_p(B)$ defined as

$$\Theta_k = \left\{ \theta \in L_2[0, 1] : \theta(x) = \sum_{j=1}^{k} \theta_j \phi_j(x), \quad \sum_{j=1}^{k} \theta_j^2 j^{2p} \leq B \right\}.$$

The parameter space can be regarded as the union of the finite-dimensional subspaces Θ_k, i.e., $\Theta_p(B) = \cup_{k=1}^{\infty} \Theta_k$, in the same way as $\mathcal{F} = \cup_{k=1}^{\infty} \mathcal{F}_k$, with $\mathcal{F}_k = \{P_\theta : \theta \in \Theta_k\}$. The sequence $\{\Theta_k\}_{k=1}^{\infty}$ can be thought of as a sieve in the sense of Grenander (1981). The basic idea is to approximate the infinite-dimensional space with a sequence of finite-dimensional subspaces. A justification for adopting this sieve could be the fact that the maximum likelihood estimator converges to f_0 at the minimax rate $n^{-2p/(2p+1)}$ under the Kullback-Leibler loss, see Barron *et al.* (1991).

When the sample size is n, let the first k Fourier coefficients to be independent and normally distributed, i.e., $\theta_j \sim N(0, j^{-(2p+1)})$, $j = 1, \ldots, k$, and let π_n be the k-variate normal density

$$\pi_n(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{j=1}^{k} \pi_j(\theta_j) = \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi j^{-(2p+1)}}} \exp\left\{ -\frac{\theta_j^2}{2j^{-(2p+1)}} \right\}. \tag{4.1}$$

The prior π_n can be truncated to have support on Θ_k. Let μ_n stand for its restriction and let Π_n stand for the corresponding prior induced on $\mathcal{F}_k$. The following assertion holds true.

Theorem 4.1. *Let Π_n be the sample-size dependent prior supported on $\mathcal{F}_k$ induced by the restriction of* (4.1) *to Θ_k. Then, for each $P_0 \in \mathcal{F}$, the conditions of Theorem 1.1 are fulfilled for $\varepsilon_n \propto n^{-p/(2p+1)}$. The posterior converges in probability at the rate $n^{-p/(2p+1)}$, relative to the Hellinger distance, uniformly on $\mathcal{F}$.*

Proof. As argued in Theorem 3.2, in order to check that $\Pi_n(M_n) \gtrsim e^{-Cnr_n^2}$, for some $C > 0$, it suffices to show that $\mu_n(\{\theta \in \Theta_k : \|\theta - \theta_0\|_2^2 < cr_n^2\}) \gtrsim e^{-Cnr_n^2}$, for a suitable constant $c > 0$. Along the lines of Shen *et al.* (2001, pages 697-698), it is seen that, by choosing $k \equiv k_n = \lceil (2B/cr_n^2)^{1/(2p)} \rceil$, since $\sum_{j=k+1}^{\infty} \theta_{0j} \leq Bk^{-2p}$, it turns out that, for constants c_1, $C > 0$,

$$\begin{aligned}
\mu_n\left(\left\{\theta \in \Theta_k : \|\theta - \theta_0\|_2^2 < cr_n^2\right\}\right) &\geq \pi_n\left(\left\{\theta : \|\theta - \theta_0\|_2^2 < cr_n^2, \quad \sum_{j=1}^{k} \theta_j^2 j^{2p} \leq B\right\}\right) \\
&= \pi_n\left(\left\{\theta : \sum_{j=1}^{k} (\theta_j - \theta_{0j})^2 + \sum_{j=k+1}^{\infty} \theta_{0j}^2 < cr_n^2, \quad \sum_{j=1}^{k} \theta_j^2 j^{2p} \leq B\right\}\right) \\
&\geq \pi_n\left(\left\{\theta : \sum_{j=1}^{k} (\theta_j - \theta_{0j})^2 < c_1 r_n^2\right\}\right) \\
&\gtrsim e^{-Cnr_n^2},
\end{aligned}$$

which concludes the proof. □

Even though the above prior is easy to work with and leads to the desired rate, it is a *sample-size dependent* prior, therefore, a prior not in accordance with the Bayesian approach in the truest sense. For this reason, a legitimate prior, whose posterior converges at the optimal rate, is preferable.

5. Convergence rate for the posterior led by a sieve prior

In this section, we show how the sieve idea can be combined with the Bayesian approach to define a legitimate prior, i.e., a prior not depending on data, called "sieve prior". The basic idea is to put a prior on the index of the sieve. The dimension of the exponential family is thus chosen by a prior on the natural numbers. The resulting prior has a hierarchical structure. If we assume that $k \sim \rho$, and conditionally on k, that π_k is a prior on the first k Fourier coefficients, then the model is

$$X_i|\theta \sim f_\theta$$

$$\theta|k \sim \pi_k$$

$$k \sim \rho.$$

If the mixing parameter k has probability $P(K = k) = \lambda_k$, then the prior can be written as

$$\pi = \sum_{k=1}^{\infty} \lambda_k \pi_k,$$

where $\{\lambda_k\}_{k=1}^{\infty}$ is a sequence such that $\lambda_k \geq 0$ and $\sum_{k=1}^{\infty} \lambda_k = 1$, and each π_k is supported on $\mathbb{R}^k$. Sieve priors have been used by Zhao (2000) and Shen *et al.* (2001) for determining rates of convergence of the Bayes's estimator and the posterior distribution, respectively, for non-parameteric regression. We assume the priors to be elicited in the following way. Let π_k be such that the Fourier coefficients are independent and, for $q = p + 1/2$,

$$\text{(5.1)} \qquad \begin{cases} \theta_j \sim N(0, j^{-2q}), & \text{if } j = 1, \ldots, k, \\ \theta_j \equiv 0, & \text{if } j > k. \end{cases}$$

For positive numbers A_0 and A, let

$$\text{(5.2)} \qquad \lambda_k \geq A_0 e^{-Ak}, \qquad k = 1, \ldots.$$

We shall prove that the posterior corresponding to the restriction of π to $\Theta_p(B)$ converges at the optimal rate uniformly on the parameter space. For any Borel set $A \subseteq L_2[0, 1]$, let as usual

$$\begin{aligned} \mu(A) &= \frac{\pi(A \cap \Theta_p(B))}{\pi(\Theta_p(B))} \\ &= \frac{\sum_{k=1}^{\infty} \lambda_k \pi_k(A \cap \Theta_k)}{\sum_{k=1}^{\infty} \lambda_k \pi_k(\Theta_k)}, \end{aligned} \qquad \text{(5.3)}$$

where the Θ_k's are as in Section 4. With obvious meaning of symbols, it is

$$\Pi(H_{\varepsilon_n}) = \frac{\sum_{k=1}^{\infty} \lambda_k \Pi_k(H_{\varepsilon_n} \cap \mathcal{F}_k)}{\sum_{k=1}^{\infty} \lambda_k \Pi_k(\mathcal{F}_k)},$$

where $\mathcal{F}_k = \{P_\theta : \theta \in \Theta_k\}$. The aforesaid result is now stated.

Theorem 5.1. *Let Π be the prior supported on $\mathcal{F}$ induced by the restriction of $\pi = \sum_{k=1}^{\infty} \lambda_k \pi_k$ to $\Theta_p(B)$, where π_k and λ_k are elicited as prescribed by (5.1) and (5.2). Then, for each $P_0 \in \mathcal{F}$, the posterior converges almost surely at the rate $n^{-p/(2p+1)}$, relative to the Hellinger distance.*

Proof. The basic arguments parallel those of the proofs of the previous theorems. The only difference is the starting point:

$$\begin{aligned}
\mu\left(\left\{\theta \in \Theta_p(B) : \|\theta - \theta_0\|_2^2 < cr_n^2\right\}\right) &\geq \pi\left(\left\{\theta \in \Theta_p(B) : \|\theta - \theta_0\|_2^2 < cr_n^2\right\}\right) \\
&= \sum_{k=1}^{\infty} \lambda_k \pi_k\left(\left\{\theta \in \Theta_p(B) : \|\theta - \theta_0\|_2^2 < cr_n^2\right\}\right) \\
&\geq \lambda_k \pi_k\left(\left\{\theta \in \Theta_p(B) : \|\theta - \theta_0\|_2^2 < cr_n^2\right\}\right) \\
&\geq A_0 e^{-Ak} \pi_k\left(\left\{\theta \in \Theta_p(B) : \|\theta - \theta_0\|_2^2 < cr_n^2\right\}\right) \\
&\geq e^{-Cnr_n^2},
\end{aligned}$$

for some $C > 0$. □

In principle, a different sieve prior yielding the same rate might be constructed as in Theorem 3.1 of Ghosal *et al.* (2000, pages 508-509). However, this construction is difficult to perform. It requires the bracketing approximations of the parameter space to be known.

Arguing as in Corollary 3.1, it can be shown that the Bayes' density estimator attains the minimax rate.

Some remarks are in order. Whichever is the prior adopted among those suggested, it seems that some kind of tail condition, more or less restrictive, is unavoidable. As for the sieve prior, recall the hypothesis on the distribution of the mixing parameter.

Note that $n^{-p/(2p+1)} \leq n^{-1/3}$ because $p \geq 1$ and $n^{-p/(2p+1)} \to n^{-1/2}$ as $p \to \infty$. This means that the smoother is the log-density, the faster is the convergence. For p big enough, the posterior converges at nearly parametric rate. However, the degree of smoothness p is unknown in general and the priors above suggested depend on it. Therefore, it would be desirable to define priors without prior knowledge of the degree of smoothness of the density function such that the posterior distributions converge at the optimal rate. Then, the Bayes' estimator would provide an adaptive procedure for density estimation. We hope to report on this issue in a future work.

References

[1] Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. *Technical Report n. 7, Dept. Statist., Univ. of Illinois, Champaign, IL.*

[2] Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, **27**, 536–561.
[3] Barron, A. R. and Sheu, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.*, **19**, 1347–1369. Correction, 2284.
[4] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, **4**, 329–375.
[5] Birman, M. S. and Solomjak, M. Z. (1967). Piecewise-polynomial approximation of functions of the classes W_p^α. *Mat. Sbornik*, **73**, 295–317.
[6] Crain, B. R. (1974). Estimation of distributions using orthogonal expansions. *Ann. Statist.*, **2**, 454–463.
[7] Crain, B. R. (1976). More on estimation of distributions using orthogonal expansions. *J. Amer. Statist. Assoc.*, **71**, 741–745.
[8] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, **28**, 500–531.
[9] Grenander, U. (1981). *Abstract Inference.* Wiley, New York.
[10] Ibragimov, I. A. and Hasminskii, R. Z. (1977). Estimation of infinite-dimensional parameter in white Gaussian noise. *Dokl. Akad. Nauk SSSR*, **236**, 1053–1056 (in Russian).
[11] Ibragimov, I. A. and Hasminskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.
[12] Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.*, **83**, 509–516.
[13] Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, **78**, 531–543.
[14] Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredachi Informatsii*, **16**, 52–68 (in Russian). (Translation in *Problems Inform. Transmission*, **16**, 120–133).
[15] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, **29**, 687–714.
[16] Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.*, **26**, 1215–1241.
[17] Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *J. R. Statist. Soc.* B, **63**, 811–821.
[18] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, **27**, 1564–1599.
[19] Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.*, **28**, 532–552.

C. Scricciolo,
Department of Statistical Sciences,
University of Padova,
Via C. Battisti, 241/243,
35121I - Padova,
Italy
E-mail address: catia@stat.unipd.it